Journal of Machine Learning Research 10 (2009) 2349-2374 Submitted 8/08; Revised 7/09; Published 10/09

Estimating Labels from Label Proportions*

Novi Quadrianto NOVI.QUAD@GMAIL.COM
Statistical Machine Learning
National ICT Australia
Locked Bag 8001
Canberra ACT 2601, Australia

Alex J. Smola ALEX@SMOLA.ORG
Yahoo! Research
4401 Great America Parkway
Santa Clara CA 95054, USA

Tibério S. Caetano TIBERIO.CAETANO@GMAIL.COM
Statistical Machine Learning
National ICT Australia
Locked Bag 8001
Canberra ACT 2601, Australia

Quoc V. Le QUOCLE@STANFORD.EDU
AI Lab
Department of Computer Science Department
Stanford University
Stanford CA 94305, USA

Editor: Tommi Jaakkola

Abstract

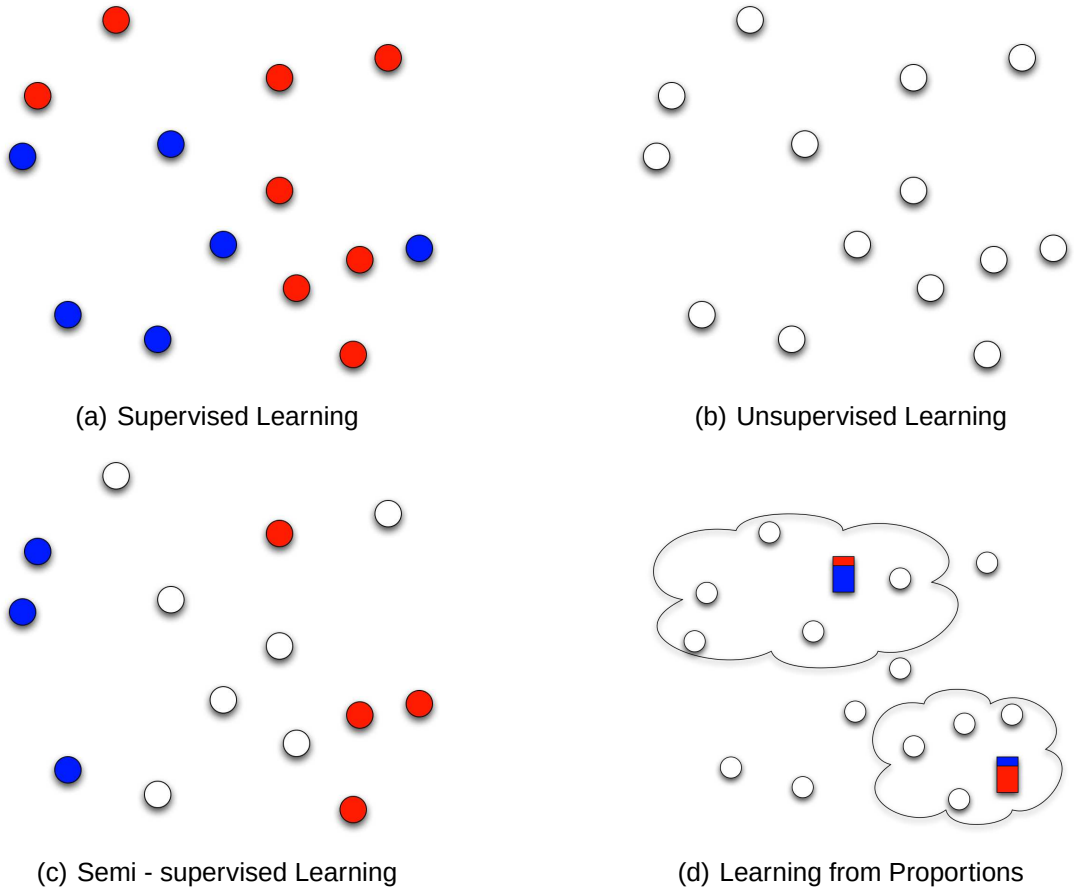
Consider the following problem: given sets of unlabeled observations, each set with known label proportions, predict the labels of another set of observations, possibly with known label proportions. This problem occurs in areas like e-commerce, politics, spam filtering and improper content detection. We present consistent estimators which can reconstruct the correct labels with high probability in a uniform convergence sense. Experiments show that our method works well in practice.

Keywords: unsupervised learning, Gaussian processes, classification and prediction, probabilistic models, missing variables

1. Introduction

Different types of learning problems assume different problem settings. In *supervised* learning, we are given sets of labeled instances. Another learning type called *unsupervised* learning focuses on the setting where unlabeled instances are given. Recently, it has been realized that unlabeled instances when used in conjunction with a small amount of labeled instances can deliver considerable learning performance improvement in comparison to using labeled instances alone. This leads to a *semi-supervised* learning setting.

∗. A short version of this paper appeared in Quadrianto et al. (2008).



(a) Supervised Learning

(b) Unsupervised Learning

(c) Semi - supervised Learning

(d) Learning from Proportions

Figure 1: Different types of learning problems (colors encode class labels). 1(a) - `supervised learning`: only labeled instances are given; 1(b) - `unsupervised learning`: only unlabeled instances are given; 1(c) - `semi-supervised learning`: both labeled and unlabeled instances are given; 1(d): `learning from proportions`: at least as many data aggregates (groups of data with their associated class label proportions) as there are number of classes are given.

We are interested in a learning setting where groups of unlabeled instances are given. The number of group is at least as many as number of classes. Each group is equipped with information on class label *proportions*. We called this informative group as aggregate (see Figure 1 for an illustration). This type of learning problem appears in areas like e-commerce, politics, spam filtering and improper content detection, as we illustrate below.

Assume that an internet services company wants to increase its profit in sales. Obviously sending out discount coupons will increase sales, but sending coupons to customers who would have purchased the goods anyway decreases the margins. Alternatively, failing to send coupons to customers who would only buy in case of a discount reduces overall sales. We would like to identify the class of would-be customers who are most likely to change their purchase decision when receiving a coupon. The problem is that there is no direct access to a sample of would-be customers.

Typically only a sample of people who buy regardless of coupons (those who bought when there was no discount) and a mixed sample (those who bought when there was discount) are available. The mixing *proportions* can be reliably estimated using random assignment to control and treatment groups. How can we use this information to determine the would-be customers?

Politicians face the same problem. They can rely on a set of always-favorable voters who will favor them regardless, plus a set of swing voters who will make their decision dependent on what the candidates offer. Since the candidate's resources (finance, ability to make election promises, campaign time) are limited, it is desirable for them to focus their attention on that part of the demographic where they can achieve the largest gains. Previous elections can directly reveal the profile of those who favor regardless, that is those who voted in favor where low campaign resources were committed. Those who voted in favor where substantial resources were committed can be either swing voters or always-favorable. So in a typical scenario there is no separate sample of swing voters.

Likewise, consider the problem of spam filtering. Data sets of spam are likely to contain almost pure spam (this is achieved e.g. by listing e-mails as spam bait), while user's inboxes typically contain a mix of spam and non-spam. We would like to use the inbox data to improve estimation of spam. In many cases it is possible to estimate the *proportions* of spam and non-spam in a user's inbox much more cheaply than the actual labels. We would like to use this information to categorize e-mails into spam and non-spam.

Similarly, consider the problem of filtering images with "improper content". Data sets of such images are readily accessible thanks to user feedback, and it is reasonable to assume that this labeling is highly reliable. However the rest of images on the web (those not labeled) is a far larger data set, albeit without labels (after all, this is what we would like to estimate the labels for). That said, it is considerably cheaper to obtain a good estimate of the *proportions* of proper and improper content in addition to having one data set of images being of likely improper content. We would like to obtain a classifier based on this information.

2. Problem Definition

In this paper, we present a method that makes use of the knowledge of label proportions *directly*. As motivated by the above examples, our method would be practically useful in many domains such as identifying potential customers, potential voters, spam e-mails and improper images. We also prove bounds indicating that the estimates obtained are close to those from a fully labeled scenario.

Before defining the problem, we emphasize that the formal setting is more general than the above examples might suggest. More specifically, we may not require *any* label to be known, only their proportions within each of the involved data sets. Also the general problem is not restricted to the binary case but instead can deal with large numbers of classes. Finally, it is possible to apply our method to problems where the *test label proportions* are unknown, too. This simple modification allows us to use this technique whenever covariate shift via label bias is present.

Formally, in a learning from proportions setting, we are given n sets of observations $X_i = \{x_1^i, \ldots, x_{m_i}^i\}$ of respective sample sizes m_i (calibration set) $i = 1, \ldots, n$ as well as a set $X = \{x_1, \ldots, x_m\}$ (test set). Moreover, we are given the fractions π_{iy} of labels $y \in \mathcal{Y}$ ($|\mathcal{Y}| \leq n$) contained in each set X_i. These fractions form a full (column) rank mixing matrix, $\pi \in \mathbb{R}^{n \times |\mathcal{Y}|}$ with the constraint that each row sums up to 1 and all entries are nonnegative. The marginal probability $p(y)$ of the test set X may or may not be known. Note that the label dictionaries $\mathcal{Y}_i$ do not need to be the same

across all sets i (define $\mathcal{Y} := \cup_i \mathcal{Y}_i$) and we also allow for $\pi_{iy} = 0$ if needed. It is our goal to design algorithms which are able to obtain conditional class probability estimates $p(y|x)$ solely based on this information.

As an illustration, take the spam filtering example. We have $X_1 =$ "mail in spam box" (only spam) and $X_2 =$ "mail in inbox" (spam mixed with non-spam). Also suppose that we may know the proportion of spam vs non-spam in our inbox is $1:9$. That means, we know: $\pi_{1,\text{spam}} = 1.0, \pi_{1,\text{non-spam}} = 0, \pi_{2,\text{spam}} = 0.1$ and $\pi_{2,\text{non-spam}} = 0.9$. The test set X then may be X_2 itself, for example. Thus, the marginal probability of the test set will simply be: $p(y = \text{spam}) = 0.1, p(y = \text{non}-\text{spam}) = 0.9$. The goal is to find $p(\text{spam}|\text{mail})$ in X. Note that, in general, our setting is different and more difficult than that of transduction. The latter requires at least some labeled instances of *all classes* are given. In the spam filtering example, we have no pure non-spam instances.

Key to our proposed solution is a conditional independence assumption, $x \perp\!\!\!\perp i \,|\, y$. In other words, we assume that the *conditional* distribution of x is independent of the index i, as long as we know the label y. This is a crucial assumption: after all, we want the distributions within each class to be independent of which aggregate they can be found in. If this were not the case it would be impossible to infer about the distribution on the test set from the (biased) distributions over the aggregates.

3. Mean Operators

Our idea relies on uniform convergence properties of the expectation operator and of corresponding risk functionals (Altun and Smola, 2006; Dudík and Schapire, 2006). In doing so, we are able to design estimators with the same performance guarantees in terms of uniform convergence as those with full access to the label information.

At the heart of our reasoning lies the fact that many estimators rely on data by solving a convex optimization problem. We begin our exposition by discussing how this strategy can be employed in the context of exponential families. Subsequently we state convergence guarantees and we discuss how our method can be extended to other estimates such as Csiszar and Bregman divergences and other function spaces.

3.1 Exponential Families

Denote by $\mathcal{X}$ the space of observations and let $\mathcal{Y}$ be the space of labels. Moreover, let $\phi(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ be a feature map into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with kernel $k((x,y),(x',y'))$. In this case we may state conditional exponential models via

$$p(y|x,\theta) = \exp\left(\langle \phi(x,y), \theta \rangle - g(\theta|x)\right) \text{ with } g(\theta|x) = \log \sum_{y \in \mathcal{Y}} \exp \langle \phi(x,y), \theta \rangle ,$$

where the normalization g is called the log-partition function, often referred to as the cumulant generating function. Note that while in general there is no need for $\mathcal{Y}$ to be discrete, we make this simplifying assumption in order to be able to reconstruct the class probabilities efficiently. For $\{(x_i,y_i)\}$ drawn iid from a distribution $p(x,y)$ on $\mathcal{X} \times \mathcal{Y}$ the conditional log-likelihood is given by

$$\log p(Y|X,\theta) = \sum_{i=1}^{m} \left[\langle \phi(x_i,y_i), \theta \rangle - g(\theta|x_i)\right] = m \langle \mu_{XY}, \theta \rangle - \sum_{i=1}^{m} g(\theta|x_i),$$

where the empirical mean in feature space μ_{XY} is defined as in Table 2. In order to avoid overfitting one commonly maximizes the log-likelihood penalized by a prior $p(\theta)$. This means that we need to solve the following optimization problem

$$\theta^* := \operatorname*{argmin}_\theta \left[-\log\{p(Y|X,\theta)p(\theta)\}\right]. \tag{1}$$

For instance, for a Gaussian prior on θ, i.e. for

$$-\log p(\theta) = \lambda \|\theta\|^2 + \text{const.},$$

we have

$$\theta^* = \operatorname*{argmin}_\theta \left[\sum_{i=1}^m g(\theta|x_i) - m\langle \mu_{XY}, \theta\rangle + \lambda\|\theta\|^2\right]. \tag{2}$$

The problem is that in our setting we do not know the labels y_i, so the sufficient statistics μ_{XY} cannot be computed exactly. Note, though that the only place where the labels enter the estimation process is via the mean μ_{XY}. Our strategy is to exploit the fact that this quantity, however, is statistically well behaved and converges under relatively mild technical conditions at rate $O(m^{-\frac{1}{2}})$ to its expected value

$$\mu_{xy} := \mathbf{E}_{(x,y)\sim p(x,y)}[\phi(x,y)],$$

as will be shown in Theorem 3. Our goal therefore will be to estimate μ_{xy} and use it as a proxy for μ_{XY}, and only then solve (2) with the estimated $\hat{\mu}_{XY}$ instead of μ_{XY}. We will discuss explicit convergence guarantees in Section 5 after describing how to compute the mean operator in detail.

3.2 Estimating the Mean Operator

In order to obtain θ^* we would need μ_{XY}, which is impossible to compute exactly, since we do not have Y. However, we know that μ_{XY} converges to μ_{xy}. Hence, if we are able to approximate μ_{xy} then this, in turn, will be a good estimate for μ_{XY}.

Our quest is therefore as follows: express μ_{xy} as a linear combination over expectations with respect to the distributions on the data sets $X_1,\ldots,X_n$ (where $n \geq |\mathcal{Y}|$). Secondly, show that the expectations of the distributions having generated the sets X_i ($\mu_X^{\text{set}}[i,y']$, see Table 2), can be approximated by empirical means ($\mu_X^{\text{set}}[i,y']$, see Table 2). Finally, we need to combine both steps to provide guarantees for μ_{XY}.

It will turn out that in certain cases some of the algebra can be sidestepped, in particular whenever we may be able to identify several sets with each other (e.g. the test set X is one of the calibration data sets X_i) or whenever $\phi(x,y)$ factorizes into $\psi(x)\otimes\varphi(y)$. We will discuss these simplifications in Section 4.

3.2.1 MEAN OPERATOR

Since μ_{xy} is a linear operator mapping $p(x,y)$ into a Hilbert Space we may expand μ_{xy} via

$$\mu_{xy} = \mathbf{E}_{(x,y)\sim p(x,y)}[\phi(x,y)] = \sum_{y\in\mathcal{Y}} p(y)\mathbf{E}_{x\sim p(x|y)}[\phi(x,y)] = \sum_{y\in\mathcal{Y}} p(y)\mu_x^{\text{class}}[y,y],$$

X_i	i^{th} set of observations: $X_i = \{x^i_1, \ldots, x^i_{m_i}\}$
m_i	number of observations in X_i
X	test set of observations: $X = \{x_1, \ldots, x_m\}$
Y	test set of labels: $Y = \{y_1, \ldots, y_m\}$
m	number of observations in the test set X
π_{iy}	proportion of label y in set i
$\phi(x,y)$	map from (x,y) to a Hilbert Space

Table 1: Notations used in the paper.

Expectations with respect to the model:

$$\begin{aligned}
\mu_{xy} &:= \mathbf{E}_{(x,y)\sim p(x,y)}[\phi(x,y)] \\
\mu_x^{\text{class}}[y,y'] &:= \mathbf{E}_{(x)\sim p(x|y)}[\phi(x,y')] \\
\mu_x^{\text{set}}[i,y'] &:= \mathbf{E}_{(x)\sim p(x|i)}[\phi(x,y')] \\
\mu_x^{\text{class}}[y] &:= \mathbf{E}_{(x)\sim p(x|y)}[\psi(x)] \\
\mu_x^{\text{set}}[i] &:= \mathbf{E}_{(x)\sim p(x|i)}[\psi(x)]
\end{aligned}$$

Expectations with respect to data:

$$\begin{aligned}
\mu_{XY} &:= \tfrac{1}{m}\textstyle\sum_{i=1}^m \phi(x_i,y_i) \\
\text{(1a)}\quad \mu_X^{\text{set}}[i,y'] &:= \tfrac{1}{m_i}\textstyle\sum_{x\in X_i} \phi(x,y') \ \textbf{(known)} \\
\text{(1b)}\quad \mu_X^{\text{set}}[i] &:= \tfrac{1}{m_i}\textstyle\sum_{x\in X_i} \psi(x) \ \textbf{(known)}
\end{aligned}$$

Estimates:

$$\begin{aligned}
\text{(2)}\quad \hat{\mu}_x^{\text{class}} &= (\pi^\top\pi)^{-1}\pi^\top\mu_X^{\text{set}} \\
\text{(3a)}\quad \hat{\mu}_{XY} &= \textstyle\sum_{y\in\mathcal{Y}} p(y)\hat{\mu}_x^{\text{class}}[y,y] \\
\text{(3b)}\quad \hat{\mu}_{XY} &= \textstyle\sum_{y\in\mathcal{Y}} p(y)\varphi(y)\otimes\hat{\mu}_x^{\text{class}}[y] \\
\text{(4)}\quad \hat{\theta}^* &\text{ solution of (2) for } \mu_{XY} = \hat{\mu}_{XY}.
\end{aligned}$$

Table 2: Major quantities of interest in the paper. Numbers on the left represent the order in which the corresponding quantity is computed in the algorithm (letters denote the variant of the algorithm: 'a' for general feature map $\phi(x,y)$ and 'b' for factorizing feature map $\phi(x,y) = \psi(x)\otimes\varphi(y)$). Lowercase subscripts refer to model expectations, uppercase subscripts are sample averages.

where the shorthand $\mu_x^{\text{class}}[y,y]$ is defined in Table 2. This means that if we were able to compute $\mu_x^{\text{class}}[y,y]$ we would be able to "reassemble" μ_{xy} from its individual components. We now show that $\mu_x^{\text{class}}[y,y]$ can be estimated directly.

Our conditional independence assumption, $p(x|y,i) = p(x|y)$, yields the following:

$$p(x|i) = \sum_y p(x|y,i)p(y|i) = \sum_y p(x|y)\pi_{iy}. \tag{3}$$

In the above equation, we form a mixing matrix π with the element $\pi_{iy} = p(y|i)$. This allows us to define the following means

$$\mu_x^{\text{set}}[i,y'] := \mathbf{E}_{x\sim p(x|i)}[\phi(x,y')] \overset{(3)}{=} \sum_y \pi_{iy}\mu_x^{\text{class}}[y,y'].$$

Algorithm 1

Input data sets X, $\{X_i\}$, probabilities π_{iy} and $p(y)$
for $i = 1$ **to** n **and** $y' \in \mathcal{Y}$ **do**
Compute empirical means $\mu_X^{\text{set}}[i,y']$
end for
Compute $\hat{\mu}_x^{\text{class}} = (\pi^\top \pi)^{-1} \pi^\top \mu_X^{\text{set}}$
Compute $\hat{\mu}_{XY} = \sum_{y \in \mathcal{Y}} p(y) \hat{\mu}_x^{\text{class}}[y,y]$
Solve the minimization problem

$$\hat{\theta}^* = \operatorname*{argmin}_\theta \left[\sum_{i=1}^{m} g(\theta | x_i) - m \langle \hat{\mu}_{XY}, \theta \rangle + \lambda \|\theta\|^2 \right]$$

Return $\hat{\theta}^*$.

Note that in order to compute $\mu_x^{\text{set}}[i,y']$ we do *not* need any label information with respect to $p(x|i)$. It is simply the expectation of $\phi(\cdot,y')$ on the distribution of bag i. However, since we have at least $|\mathcal{Y}|$ of those equations and we assumed that π has full column rank, they allow us to solve a linear system of equations and compute $\mu_x^{\text{class}}[y,y]$ from $\mu_x^{\text{set}}[i,y']$ for all i. In shorthand we may use

$$\mu_x^{\text{set}} = \pi \mu_x^{\text{class}} \text{ and hence } \mu_x^{\text{class}} = (\pi^\top \pi)^{-1} \pi^\top \mu_x^{\text{set}} \tag{4}$$

to compute $\mu_x^{\text{class}}[y,y]$ for all $y \in \mathcal{Y}$. With some slight abuse of notation we have μ_x^{class} and μ_x^{set} represent the *matrices* of terms $\mu_x^{\text{class}}[y,y']$ and $\mu_x^{\text{set}}[i,y']$ respectively. There will be as many matrices as the dimensions of $\phi(x,y)$, thus (4) has to be solved separately for each dimension of $\phi(x,y)$.

Obviously we cannot compute $\mu_x^{\text{set}}[i,y']$ explicitly, since we only have *samples* from $p(x|i)$. However the same convergence results governing the convergence of μ_{XY} to μ_{xy} also hold for the convergence of $\mu_X^{\text{set}}[i,y']$ to $\mu_x^{\text{set}}[i,y']$. Hence we may use the empirical average $\mu_X^{\text{set}}[i,y']$ as the estimate for $\mu_x^{\text{set}}[i,y']$ and from that find an estimate for μ_{XY}.

3.2.2 BIG PICTURE

Overall, our strategy is as follows: use empirical means on the bags X_i to approximate expectations with respect to the bag distribution. Use the latter to compute expectations with respect to a given label, and finally, use the means conditional on the label distribution to obtain μ_{xy} which is a good proxy for μ_{XY} (see Algorithm 1), i.e.

$$\mu_X^{\text{set}}[i,y'] \longrightarrow \mu_x^{\text{set}}[i,y'] \longrightarrow \mu_x^{\text{class}}[y,y'] \longrightarrow \mu_{xy} \longrightarrow \mu_{XY}.$$

For the first and last step in the chain we can invoke uniform convergence results. The remaining two steps in the chain follow from linear algebra. As we shall see, whenever there are considerably more bags than classes we can exploit the overdetermined system to our advantage to reduce the overall estimation error and use a rescaled version of (4).

4. Special Cases

In some cases the calculations described in Algorithm 1 can be carried out more efficiently. They arise whenever the matrix π has special structure or whenever the test set and one of the training

sets coincide. Moreover, we may encounter situations where the fractions of observations in the test set are unknown and we would like, nonetheless, to find a good proxy for μ_{XY}.

4.1 Minimal Number of Sets

Assuming that $|\mathcal{Y}| = n$ and that π has full rank it follows that $(\pi^\top\pi)^{-1}\pi^\top = \pi^{-1}$. Hence we can obtain the proxy for μ_{XY} more directly via $\mu_x^{\text{class}} = \pi^{-1}\mu_x^{\text{set}}$.

4.2 Testing on One of the Calibration Sets

Note that there is no need for requiring that the test set X be different from one of the calibration sets (vide example in Problem Definition). In particular, when $X = X_i$ the uncertainty in the estimate of μ_{XY} can be greatly reduced provided that the estimate of μ_{XY} as given in (4) contains a large fraction of the mean of at least one of the classes. We will discuss this situation in more detail when it comes to binary classification since there the advantages will be most obvious.

4.3 Special Feature Map

Whenever the feature map $\phi(x,y)$ factorizes into $\psi(x) \otimes \varphi(y)$ we can simplify calculation of the means considerably. More specifically, instead of estimating $O(|\mathcal{Y}| \cdot n)$ parameters we only require calculation of $O(n)$ terms. The reason for this is that we may pull the dependency on y out of the expectations. Defining $\mu_x^{\text{class}}[y], \mu_x^{\text{set}}[i]$, and $\mu_X^{\text{set}}[i]$ as in Table 2 allows us to simplify

$$\hat{\mu}_{XY} = \sum_{y\in\mathcal{Y}} p(y)\varphi(y) \otimes \hat{\mu}_x^{\text{class}}[y] \text{ where } \hat{\mu}_x^{\text{class}} = (\pi^\top\pi)^{-1}\pi^\top\mu_X^{\text{set}}. \tag{5}$$

Here the last equation is understood to apply to the vector of means $\mu_x := (\mu[1], \ldots, \mu[n])$ and μ_X accordingly. A significant advantage of (5) is that we only need to perform $O(n)$ averaging operations rather than $O(n \cdot |\mathcal{Y}|)$. Obviously the cost of computing $(\pi^\top\pi)^{-1}\pi^\top$ remains unchanged but the latter is negligible compared to the operations in Hilbert Space. Note that $\psi(x) \in \mathbb{R}^D$ denotes an arbitrary feature representation of the inputs, which in many cases can be defined implicitly via a kernel function. As the joint feature map $\phi(x,y)$ factorizes into $\psi(x) \otimes \varphi(y)$, we can write the inner product in the joint representation as $\langle\phi(x,y),\phi(x',y')\rangle = \langle\psi(x),\psi(x')\rangle\langle\varphi(y),\varphi(y')\rangle = k(x,x')k(y,y')$. In general, the kernel function on inputs and labels can be different. Specifically, for a label diagonal kernel $k(y,y') = \delta(y,y')$, the standard winner-takes-all multiclass classification is recovered (Tsochantaridis et al., 2005). With this setting, the input feature $\psi(x)$ can be defined implicitly via a kernel function by invoking the Representer Theorem (Schölkopf and Smola, 2002).

4.4 Binary Classification

One may show (Hofmann et al., 2006) that the feature map $\phi(x,y)$ takes on a particularly appealing form of $\phi(x,y) = y\psi(x)$ where $y \in \{\pm 1\}$. This follows since we can always re-calibrate $\langle\phi(x,y),\theta\rangle$ by an offset independent of y such that $\phi(x,1) + \phi(x,-1) = 0$.

If we moreover assume that X_1 only contains class 1 and $X_2 = X$ contains a mixture of classes with labels 1 and -1 with proportions $p(1) =: \rho$ and $p(-1) = 1-\rho$ respectively, we obtain the mixing matrix

$$\pi = \begin{bmatrix} 1 & 0 \\ \rho & 1-\rho \end{bmatrix} \Rightarrow \pi^{-1} = \begin{bmatrix} 1 & 0 \\ \frac{-\rho}{1-\rho} & \frac{1}{1-\rho} \end{bmatrix}.$$

Plugging this into (5) yields

$$\begin{aligned}\hat{\mu}_{XY} &= \rho\mu_X^{\text{set}}[1] - (1-\rho)\left[\tfrac{-\rho}{1-\rho}\mu_X^{\text{set}}[1] + \tfrac{1}{1-\rho}\mu_X^{\text{set}}[2]\right] \\ &= 2\rho\mu_X^{\text{set}}[1] - \mu_X^{\text{set}}[2]. \end{aligned} \tag{6}$$

Consequently, taking a simple weighted difference between the averages on two sets, e.g. one set containing spam whereas the other one containing an unlabeled mix of spam and non-spam, allows one to obtain the sufficient statistics needed for estimation.

4.5 Overdetermined Systems

Assume that we have significantly more bags n than class labels $|\mathcal{Y}|$, possibly with varying numbers of observations m_i per bag. In this case it would make sense to find a weighting of the bags such that those which are largest and most relevant for the test set are given the highest degree of importance. Instead of stating the problem as one of solving a linear system we now restate it as one of solving an approximation problem. To simplify notation we assume that the feature map factorizes, i.e. that $\phi(x,y) = \psi(x) \otimes \varphi(y)$. A weighted linear combination of the squared discrepancy between the class means and the set means is given by

$$\underset{\mu_x^{\text{class}}}{\text{minimize}} \sum_{i=1}^{n} w_i \left\| \mu_X^{\text{set}}[i] - \sum_{y\in\mathcal{Y}} \pi_{iy}\mu_x^{\text{class}}[y] \right\|^2, \tag{7}$$

where w_i are some previously chosen weights which reflect the importance of each bag. Typically we might choose $w_i = O(m_i^{-\frac{1}{2}})$ to reflect the fact that convergence between empirical means and expectations scales with $O(m^{-\frac{1}{2}})$. Before we discuss specific methods for choosing a weighting, let us review the statistical properties of the estimator.

Remark 1 (Underdetermined Systems) *Similarly, when we have less bags n than class labels $|\mathcal{Y}|$, we can state the problem as one of solving a regularized least squares problem as follows*

$$\underset{\mu_x^{\text{class}}}{\text{minimize}} \sum_{i=1}^{n} \left\| \mu_X^{\text{set}}[i] - \sum_{y\in\mathcal{Y}} \pi_{iy}\mu_x^{\text{class}}[y] \right\|^2 + \lambda\Omega(\mu_x^{\text{class}}[y]\,\forall y \in \mathcal{Y}).$$

For example, we can let $\Omega(\mu_x^{\text{class}}[y]\,\forall y \in \mathcal{Y}) = \sum_{y\in\mathcal{Y}} \left\|\mu_x^{\text{class}}[y] - \mu_x^{\text{class}}[y+1]\right\|^2$. This makes sense whenever different labels have related means $\mu_x^{\text{class}}[y]$.

5. Convergence Bounds

The obvious question is how well $\hat{\mu}_{XY}$ manages to approximate μ_{XY} and secondly, how badly any error in estimating μ_{XY} would affect the overall quality of the solution. We approach this problem as follows: first we state the uniform convergence properties of μ_{XY} and similar empirical operators relative to μ_{xy}. Secondly, we apply those bounds to the cases discussed above, and thirdly, we show that the approximate minimizer of the log-posterior has a bounded deviation from what we would have obtained by knowing μ_{XY} exactly. Much of the reasoning follows the ideas of Altun and Smola (2006).

5.1 Uniform Convergence for Mean Operators

An important tool in studying uniform convergence properties of random variables are Rademacher averages (Ledoux and Talagrand, 1991; Mendelson, 2002). They are needed to state the key results in our context.

Definition 2 (Rademacher Averages) *Let $\mathcal{X}$ be a domain and p a distribution on $\mathcal{X}$ and assume that $X := \{x_1, \ldots, x_m\}$ is drawn iid from p. Moreover, let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. Furthermore denote by σ_i Rademacher random variables, i.e. $\{\pm 1\}$ valued with zero mean. The Rademacher average is*

$$R_m(\mathcal{F}, p) := \mathbf{E}_X \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right| \right].$$

This quantity measures the flexibility of the function class $\mathcal{F}$—in our case linear functions in $\phi(x,y)$. Altun and Smola (2006) state the following result:

Theorem 3 (Convergence of Empirical Means) *Denote by $\phi : \mathcal{X} \to \mathcal{B}$ a map into a Banach space $\mathcal{B}$, denote by $\mathcal{B}^*$ its dual space and let $\mathcal{F}$ the class of linear functions on $\mathcal{B}$ with bounded $\mathcal{B}^*$ norm by 1. Let $R > 0$ such that for all $f \in \mathcal{F}$ we have $|f(x)| \leq R$. Moreover, assume that X is an m-sample drawn from p on $\mathcal{X}$. For $\bar{\varepsilon} > 0$ we have that with probability at least $1 - \exp(-\bar{\varepsilon}^2 m / 2R^2)$ the following holds:*

$$\|\mu_X - \mu_x\|_{\mathcal{B}} \leq 2R_m(\mathcal{F}, p) + \bar{\varepsilon}.$$

For $k \geq 0$ we only have a failure probability of $1 - \exp(-\bar{\varepsilon}^2 m / R^2)$.

Theorem 4 (Bartlett and Mendelson 2002) *Whenever $\mathcal{B}$ is a Reproducing Kernel Hilbert Space with kernel $k(x,x')$ the Rademacher average can be bounded from above by $R_m(\mathcal{F}) \leq m^{-\frac{1}{2}} \left[\mathbf{E}_x[k(x,x)]\right]^{\frac{1}{2}}$.*

Our approximation error can be bounded as follows. From the triangle inequality we have:

$$\|\hat{\mu}_{XY} - \mu_{XY}\| \leq \|\hat{\mu}_{XY} - \mu_{xy}\| + \|\mu_{xy} - \mu_{XY}\|.$$

For the second term we may employ Theorem 3 directly. To bound the first term note that by linearity

$$\varepsilon := \hat{\mu}_{XY} - \mu_{xy} = \sum_y p(y) \left[(\pi^\top \pi)^{-1} \pi^\top \hat{\varepsilon} \right]_{y,y}, \tag{8}$$

where we define the "matrix" of coefficients

$$\hat{\varepsilon}\left[i, y'\right] := \mu_x^{\text{set}}[i, y'] - \mu_X^{\text{set}}[i, y']. \tag{9}$$

In the more general case of overdetermined systems we have

$$\varepsilon = \sum_y p(y) \left[(\pi^\top W \pi)^{-1} \pi^\top W \hat{\varepsilon} \right]_{y,y}.$$

Now note that all $\hat{\varepsilon}[i,y']$ also satisfy the conditions of Theorem 3 since the sets X_i are drawn iid from the distributions $p(x|i)$ respectively. We may bound each term individually in this fashion and subsequently apply the union bound to ensure that all $n \cdot |\mathcal{Y}|$ components satisfy the constraints. Hence each of the terms needs to satisfy the constraint with probability $1-\delta/(n|\mathcal{Y}|)$ to obtain an overall bound with probability $1-\delta$. To obtain bounds we would need to bound the linear operator mapping $\hat{\varepsilon}$ into ε.

Note that this statement can be improved since all errors $\hat{\varepsilon}[i,y']$ and $\hat{\varepsilon}[j,y']$ for $i \neq j$ are independent of each other simply by the fact that each bag X_i was sampled independently from the other. We will discuss this in the context of choosing a practically useful value of W below.

5.2 Special Cases

A closed form solution in the general case is not particularly useful since it depends heavily on the kernel k, the mixing proportions π and the class probabilities on the test set. However, for a number of special cases it is possible to provide more detailed explicit analysis: firstly the situation where $\phi(x,y) = \psi(x) \otimes \varphi(y)$ and secondly, the binary classification setting where $\phi(x,y) = y\psi(x)$ and $X_2 = X$, where much tighter bounds are available.

5.3 Special Feature Map with Full Rank

Here we only need to deal with n rather than with $n \times |\mathcal{Y}|$ empirical estimates, i.e. $\mu_X^{\text{set}}[i]$ vs. $\mu_X^{\text{set}}[i,y']$. Hence (8) and (9) specialize to

$$\begin{aligned}\varepsilon &= \sum_y p(y) \sum_{i=1}^n \varphi(y) \otimes \left[(\pi^\top \pi)^{-1}\pi^\top\right]_{yi} \hat{\varepsilon}[i] \\ \hat{\varepsilon}[i] &:= \mu_x^{\text{set}}[i] - \mu_X^{\text{set}}[i].\end{aligned}$$

Assume that with high probability each $\hat{\varepsilon}[i]$ satisfies $\|\hat{\varepsilon}[i]\| \leq c_i$ (we will deal with the explicit constants c_i later). Moreover, assume for simplicity that $|\mathcal{Y}| = n$ and that π has full rank (otherwise we need to follow through on our expansion using $(\pi^\top\pi)^{-1}\pi^\top$ instead of π^{-1}). This implies that

$$\begin{aligned}\|\varepsilon\|^2 &= \sum_{i,j} \langle \hat{\varepsilon}[i], \hat{\varepsilon}[j] \rangle \times \sum_{y,y'} p(y)p(y')k(y,y') \left[\pi^{-1}\right]_{yi} \left[\pi^{-1}\right]_{y'j} \\ &\leq \sum_{i,j} c_i c_j \left| \left[\pi^{-1}\right]^\top K^{y,p} \pi^{-1} \right|_{ij}, \end{aligned} \tag{10}$$

where $K^{y,p}_{y,y'} = k(y,y')p(y)p(y')$. Combining several bounds we have the following theorem:

Theorem 5 *Assume that we have n sets of observations X_i of size m_i, each of which drawn from distributions with probabilities π_{iy} of observing data with label y. Moreover, assume that $k((x,y),(x',y')) = k(x,x')k(y,y') \geq 0$ where $k(x,x) \leq 1$ and $k(y,y) \leq 1$. Finally, assume that $m = |X|$. In this case the mean operator μ_{XY} can be estimated by $\hat{\mu}_{XY}$ with probability at least $1-\delta$ with precision*

$$\|\mu_{XY} - \hat{\mu}_{XY}\| \leq \left[2 + \sqrt{\log((n+1)/\delta)}\right] \times \left[m^{-\frac{1}{2}} + \left[\sum_{i,j} m_i^{-\frac{1}{2}} m_j^{-\frac{1}{2}} \left| \left[\pi^{-1}\right]^\top K^{y,p} \pi^{-1} \right|_{ij}\right]^{\frac{1}{2}}\right].$$

Proof We begin our argument by noting that both for $\phi(x,y)$ and for $\psi(x)$ the corresponding Rademacher averages R_m for functions of RKHS norm bounded by 1 is bounded by $m^{-\frac{1}{2}}$. This is a consequence of all kernels being bounded by 1 in Theorem 4 and $k \geq 0$.

Next note that in Theorem 3 we may set $R = 1$, since for $\|f\| \leq 1$ and $k((x,y),(x,y)) \leq 1$ and $k(x,x) \leq 1$ it follows from the Cauchy Schwartz inequality that $|f(x)| \leq 1$. Solving $\delta \leq \exp -m\varepsilon^2$ for ε yields $\varepsilon \leq m^{-\frac{1}{2}} \left[2 + \sqrt{\log(1/\delta)}\right]$.

Finally, note that we have $n+1$ deviations which we need to bound: one between μ_{XY} and μ_{xy}, and n for each of the $\varepsilon[i]$ respectively. Dividing the failure probability δ into $n+1$ cases yields bounds of the form $m^{-\frac{1}{2}} \left[2 + \sqrt{\log((n+1)/\delta)}\right]$ and $m_i^{-\frac{1}{2}} \left[2 + \sqrt{\log((n+1)/\delta)}\right]$ respectively. Plugging all error terms into (10) and summing over terms yields the claim and substituting this back into the triangle inequality proves the claim. ■

5.4 Binary Classification

Next we consider the special case of binary classification where $X_2 = X$. Using (6) we see that the corresponding estimator is given by

$$\hat{\mu}_{XY} = 2\rho\mu_X^{\text{set}}[1] - \mu_X^{\text{set}}[2].$$

Since $\hat{\mu}_{XY}$ shares a significant fraction of terms with μ_{XY} we are able to obtain tighter bounds as follows:

Theorem 6 *With probability $1-\delta$ (for $1 > \delta > 0$) the following bound holds:*

$$\|\hat{\mu}_{XY} - \mu_{XY}\| \leq 2\rho \left[2 + \sqrt{\log(2/\delta)}\right] \left[m_1^{-\frac{1}{2}} + m_+^{-\frac{1}{2}}\right],$$

where m_+ is the number of observations with $y = 1$ in X_2.

Proof Denote by $\mu[X_+]$ and $\mu[X_-]$ the averages over the subsets of X_2 with positive and negative labels respectively. By construction we have that

$$\begin{aligned}\mu_{XY} &= \rho\mu[X_+] - (1-\rho)\mu[X_-] \\ \hat{\mu}_{XY} &= 2\rho\mu_X^{\text{set}}[1] - \rho\mu[X_+] - (1-\rho)\mu[X_-].\end{aligned}$$

Taking the difference yields $2\rho\left[\mu_X^{\text{set}}[1] - \mu[X_+]\right]$. To prove the claim note that we may use Theorem 3 both for $\left\|\mu_X^{\text{set}}[1] - \mathbf{E}_{x\sim p(x|y=1)}[\psi(x)]\right\|$ and for $\left\|\mu[X_+] - \mathbf{E}_{x\sim p(x|y=1)}[\psi(x)]\right\|$. Taking the union bound and summing over terms proves the claim. ■

The bounds we provided show that $\hat{\mu}_{XY}$ converges at the same rate to μ_{xy} as μ_{XY} does, assuming that the sizes of the sets X_i increase at the same rate as X.

5.5 Overdetermined Systems

Given the optimal value of weighting W, the class mean can be reconstructed as a solution of a weighted least square problem in (7) and this minimizer is given by

$$\hat{\mu}_x^{\text{class}} = (\pi^\top W \pi)^{-1} \pi^\top W \mu_X^{\text{set}} \text{ where } W = \text{diag}(w_1, \ldots, w_n) \text{ and } w_i > 0.$$

It is easy to see that whenever $n = |\mathcal{Y}|$ and π has full rank there is only one possible solution regardless of the choice of W. For overdetermined systems the choice of W may greatly affect the

quality of the solution and it is therefore desirable to choose a weighting which minimizes the error in estimating μ_{XY}.

In choosing a weighting, we may take advantage of the fact that the errors $\hat{\varepsilon}[i]$ are independent for all i. This follows from the fact that all bags are drawn independently of each other. Moreover, we know that $\mathbf{E}[\hat{\varepsilon}[i]] = 0$ for all i. Finally we make the assumption that $k(y,y') = \delta(y,y')$, that is, that the kernel in the labels is diagonal. In this situation our analysis is greatly simplified and we have:

$$\varepsilon = \sum_y \varphi(y) \otimes p(y)(\pi^\top W \pi)^{-1} \pi W \hat{\varepsilon}$$

$$\text{and hence } \mathbf{E}\left[\|\varepsilon\|^2\right] = \sum_{i=1}^n \sum_y \mathbf{E}\left[\|\hat{\varepsilon}[i]\|^2\right] W_{ii}^2 \left[\pi_i^\top (\pi^\top W \pi)^{-1}\right]_y^2 p^2(y).$$

Using the assumption that $\mathbf{E}\left[\|\hat{\varepsilon}[i]\|^2\right] = O(m_i^{-1})$ we may find a suitable scale of the weight vectors by minimizing

$$\sum_{i=1}^n \sum_y \frac{W_{ii}^2}{m_i} \left[\pi_i^\top (\pi^\top W \pi)^{-1}\right]_y^2 p^2(y) \tag{11}$$

with respect to the diagonal matrix W. Note that the optimal value of W depends *both* on the mixtures of the bags π_i *and* on the propensity of each class $p(y)$. That is, being able to well estimate a class which hardly occurs at all is of limited value.

5.6 Stability Bounds

To complete our reasoning we need to show that our bounds translate into guarantees in terms of the minimizer of the log-posterior. In other words, estimates using the correct mean μ_{XY} vs. its estimate $\hat{\mu}_{XY}$ do not differ by a significant amount. For this purpose we make use of Altun and Smola (2006, Lemma 17).

Lemma 7 *Denote by f a convex function on $\mathcal{H}$ and let $\mu, \hat{\mu} \in \mathcal{H}$. Moreover let $\lambda > 0$. Finally denote by $\theta^*, \in \mathcal{H}$ the minimizer of*

$$L(\theta, \mu) := f(\theta) - \langle \mu, \theta \rangle + \lambda \|\theta\|^2$$

with respect to θ and $\hat{\theta}^*$ the minimizer of $L(\hat{\theta}, \hat{\mu})$ respectively. In this case the following inequality holds:

$$\left\|\theta^* - \hat{\theta}^*\right\| \leq \lambda^{-1} \|\mu - \hat{\mu}\|. \tag{12}$$

This means that a good estimate for μ immediately translates into a good estimate for the minimizer of the approximate log-posterior. This leads to the following bound on the risk minimizer.

Corollary 8 *The deviation between θ^*, as defined in (1) and $\hat{\theta}^*$, the minimizer of the approximate log-posterior using $\hat{\mu}_{XY}$ rather than μ_{XY}, is bounded by $O(m^{-\frac{1}{2}} + \sum_i m_i^{-\frac{1}{2}})$.*

Finally, we may use Altun and Smola (2006, Theorem 16) to obtain bounds on the quality of $\hat{\theta}^*$ when considering how well it minimizes the *true* negative log-posterior. Using the bound

$$L(\hat{\theta}^*, \mu) - L(\theta^*, \mu) \leq \left\|\hat{\theta}^* - \theta^*\right\| \|\hat{\mu} - \mu\|$$

yields the following bound for the log-posterior:

Corollary 9 *The minimizer $\hat{\theta}^*$ of the approximate log-posterior using $\hat{\mu}_{XY}$ rather than μ_{XY} incurs a penalty of at most $\lambda^{-1}\|\hat{\mu}_{XY}-\mu_{XY}\|^2$.*

5.7 Stability Bounds under Perturbation

Denote $1 \in \{1\}^{|\mathcal{Y}|}$ as the vector of all ones and $0 \in \{0\}^{|\mathcal{Y}|}$ as the vector of all zeros. Let Δ be the perturbation matrix such that the perturbed mixing matrix $\tilde{\pi}$ is related to the original mixing matrix π by $\tilde{\pi} = \pi + \Delta$. Note that the perturbed mixing matrix $\tilde{\pi}$ still needs to have non-negative entries and each row sums up to 1, $\tilde{\pi}1 = 1$. The stochasticity constraint on the perturbed mixing matrix imposes special structure on the perturbation matrix, i.e. each row of perturbation matrix must sum up to 0, $\Delta 1 = 0$. Let $\hat{\theta}^*$ be the minimizer of (2) with mean $\hat{\mu}_{XY}$ approximated via mixing matrix π. Similarly, define $\tilde{\theta}^*$ for $\tilde{\mu}_{XY}$ with mixing matrix $\tilde{\pi}$. We would like to bound the distance $\left\|\hat{\theta}^* - \tilde{\theta}^*\right\|$ between the minimizers. Our perturbation bound relies on Lemma 7 and on the fact that we can bound the errors made in computing an (pseudo-) inverse of a matrix:

Lemma 10 (Stability of Inverses) *For any matrix norm $\|.\|$ and full rank matrices π and $\pi+\Delta$, the error between the inverses of π and $\pi+\Delta$ is bounded by*

$$\left\|\pi^{-1} - (\pi+\Delta)^{-1}\right\| \leq \left\|\pi^{-1}\right\| \left\|(\pi+\Delta)^{-1}\right\| \|\Delta\|.$$

Proof We use the following identity $\pi^{-1} - (\pi+\Delta)^{-1} = (\pi+\Delta)^{-1}\Delta\pi^{-1}$. The identity can be shown by left multiplying both sides of equation with $(\pi+\Delta)$. Finally, by submultiplicative property of a matrix norm, the inequality $\left\|\pi^{-1}\Delta(\pi+\Delta)^{-1}\right\| \leq \left\|\pi^{-1}\right\| \|\Delta\| \left\|(\pi+\Delta)^{-1}\right\|$ follows. ■

Theorem 11 (Stability of Pseudo-Inverses: Wedin 1973) *For any unitarily invariant matrix norm $\|.\|$ and full column rank matrices π and $\pi+\Delta$, the error between the pseudo-inverses of π and $\pi+\Delta$ is bounded by*

$$\left\|\pi^{\dagger} - (\pi+\Delta)^{\dagger}\right\| \leq \mu \left\|\pi^{\dagger}\right\|_{\sigma\infty} \left\|(\pi+\Delta)^{\dagger}\right\|_{\sigma\infty} \|\Delta\|,$$

where μ denotes a scalar constant depending on the matrix norm, $\|.\|_{\sigma\infty}$ denotes the spectral norm of a matrix, and the pseudo-inverse $\pi^{\dagger}$ defined as $\pi^{\dagger} := (\pi^{\top}\pi)^{-1}\pi^{\top}$.

Proof See Wedin (1973, Theorem 4.1) for a proof. ■

Remark 12 *For full rank matrices, the constant term μ in Theorem 11 is equal to unity regardless of the matrix norm considered (Wedin, 1973).*

First, we would like to bound the difference between $\hat{\mu}_{XY}$ and $\tilde{\mu}_{XY}$, i.e. $\varepsilon_p := \hat{\mu}_{XY} - \tilde{\mu}_{XY}$. For the special feature map with full rank, this translates to

$$\varepsilon_p = \sum_y p(y) \sum_{i=1}^{n} \varphi(y) \otimes \left[\pi^{-1} - \tilde{\pi}^{-1}\right]_{yi} \mu_X^{\text{set}}[i]$$

$$\|\varepsilon_p\|^2 = \sum_{i,j} \left\langle \mu_X^{\text{set}}[i], \mu_X^{\text{set}}[j] \right\rangle \times \left[(\pi^{-1} - \tilde{\pi}^{-1})^{\top} K^{y,p} (\pi^{-1} - \tilde{\pi}^{-1})\right]_{ij}.$$

Lemma 13 *Define $K^{y,p} := V_{y,p}^\top V_{y,p}$. With the spectral norm $\|.\|_{\sigma\infty}$ and a full rank mixing matrix π, the following bound holds:*

$$\|\hat{\mu}_{XY} - \tilde{\mu}_{XY}\|_{\sigma\infty} \leq \|V_{y,p}\|_{\sigma\infty} \left\|\pi^{-1}\right\|_{\sigma\infty} \|\Delta\|_{\sigma\infty} \left\|(\pi+\Delta)^{-1}\right\|_{\sigma\infty} \left[\sum_{i,j} \left\langle \mu_X^{set}[i], \mu_X^{set}[j] \right\rangle\right]^{\frac{1}{2}}. \quad (13)$$

Proof We first upper bound $\left[(\pi^{-1} - \tilde{\pi}^{-1})^\top K^{y,p} (\pi^{-1} - \tilde{\pi}^{-1})\right]_{ij}$ by $\left\|(\pi^{-1} - \tilde{\pi}^{-1})^\top K^{y,p} (\pi^{-1} - \tilde{\pi}^{-1})\right\|_{\sigma\infty}$. We factorize $K^{y,p}$ as $V_{y,p}^\top V_{y,p}$ since $K^{y,p}$ is a positive (semi-) definite matrix. The element $K^{y,p}_{y,y'} = k(y,y')p(y)p(y')$ is obtained by multiplying a kernel $k(y,y')$ with a rank-one kernel $k'(y,y') = p(y)p(y')$ where p is a positive function. This conformal transformation preserves the positive (semi-) definiteness of $K^{y,p}$ (Schölkopf and Smola, 2002). Thus, $\left\|(\pi^{-1} - \tilde{\pi}^{-1})^\top K^{y,p} (\pi^{-1} - \tilde{\pi}^{-1})\right\|_{\sigma\infty} \leq \left\|V_{y,p}(\pi^{-1} - \tilde{\pi}^{-1})\right\|_{\sigma\infty}^2 \leq$ $\left[\|V_{y,p}\|_{\sigma\infty} \left\|(\pi^{-1} - \tilde{\pi}^{-1})\right\|_{\sigma\infty}\right]^2 \leq \left[\|V_{y,p}\|_{\sigma\infty} \left\|\pi^{-1}\right\|_{\sigma\infty} \|\Delta\|_{\sigma\infty} \left\|(\pi+\Delta)^{-1}\right\|_{\sigma\infty}\right]^2$. The last inequality follows directly from Lemma 10. ■

Corollary 14 *Define $K^{y,p} := V_{y,p}^\top V_{y,p}$. With the spectral norm $\|.\|_{\sigma\infty}$ and a full column rank mixing matrix π, the following bound holds:*

$$\|\hat{\mu}_{XY} - \tilde{\mu}_{XY}\|_{\sigma\infty} \leq \sqrt{2}\,\|V_{y,p}\|_{\sigma\infty} \left\|\pi^{\dagger}\right\|_{\sigma\infty} \|\Delta\|_{\sigma\infty} \left\|(\pi+\Delta)^{\dagger}\right\|_{\sigma\infty} \left[\sum_{i,j} \left\langle \mu_X^{set}[i], \mu_X^{set}[j] \right\rangle\right]^{\frac{1}{2}}. \quad (14)$$

Proof Similar to Lemma 13 with the constant factor μ in Theorem 11 equals to $\sqrt{2}$ for a spectral norm. ■

Combining Lemma 13 for the full rank mixing matrix case (or Corollary 14 for the full column rank mixing matrix case) with Lemma 7, we are ready to state the stability bound under perturbation:

Lemma 15 (Stability Bound under Perturbation) *The distance ε_s between the two minimizers, $\hat{\theta}^*$ and $\tilde{\theta}^*$, is bounded by*

$$\varepsilon_s \leq \lambda^{-1} \|\hat{\mu}_{XY} - \tilde{\mu}_{XY}\|.$$

It is clear from (13) and (14) that the stability of our algorithm under perturbation will depend on the size of the perturbation and on the behavior of the (pseudo-) inverse of the perturbed mixing matrix. Note that by the triangle inequality, the distance in (12) can be decomposed as $\left\|\theta^* - \hat{\theta}^*\right\| \leq \left\|\theta^* - \tilde{\theta}^*\right\| + \left\|\tilde{\theta}^* - \hat{\theta}^*\right\|$ and the second term in RHS vanishes whenever the size of perturbation Δ is zero.

6. Extensions

We describe two types of extensions on our proposed estimator: function spaces and unknown label proportions on the test sets. We will discuss both of them in turn.

6.1 Function Spaces

Note that our analysis so far focused on a specific setting, namely maximum-a-posteriori analysis in exponential families. While this is a common and popular setting, the derivations are by no means restricted to this. We have the entire class of (conditional) models described by Altun and Smola (2006) and Dudík and Schapire (2006) at our disposition. They are characterized via

$$\underset{p}{\text{minimize}} -H(p) \text{ subject to } \|\mathbf{E}_{z\sim p}[\phi(z)] - \mu\| \leq \varepsilon.$$

Here p is a distribution, H is an entropy-like quantity defined on the space of distributions, and $\phi(z)$ is some evaluation map into a Banach space. This means that the optimization problem can be viewed as an approximate maximum entropy estimation problem, where we do not enforce exact moment matching of μ but rather allow ε slack. In both Altun and Smola (2006) and Dudík and Schapire (2006) the emphasis lay on *unconditional* density models: the dual of the above optimization problem. In particular, it follows that for H being the Shannon-Boltzmann entropy, the dual optimization problem is the maximum a posteriori estimation problem, which is what we are solving here.

In the conditional case, p denotes the collection of probabilities $p(y|x_i)$ and the operator $\mathbf{E}_{z\sim p}[\phi(z)] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}_{y|p(y|x_i)}[\phi(x_i,y)]$ is the conditional expectation operator on the set of observations. Finally, $\mu = \frac{1}{m}\sum_{i=1}^{m}\phi(x_i,y_i)$, that is, it describes the empirical observations. We have two design parameters:

6.1.1 FUNCTION SPACE

Depending on which Banach Space norm we may choose to measure the deviation between μ and its expectation with respect to p in terms of e.g. the ℓ_2 norm, the ℓ_1 norm or the ℓ_∞ norm. The latter leads to sparse coding and convex combinations. This means that instead of solving an optimization problem of the form of (2) we would minimize expression of the form

$$\sum_{i=1}^{m} g(\theta|x_i) - m\langle\mu_{XY},\theta\rangle + \lambda\|\theta\|_{\mathcal{B}^*}^p,$$

where $p \geq 1$ and $\mathcal{B}^*$ is the Banach space of the natural parameter θ which is dual to the space $\mathcal{B}$ associated with the evaluation functionals $\phi(x,y)$. The most popular choice for $\mathcal{B}^*$ is ℓ_1 which leads to sparse coding (Candes and Tao, 2005; Chen et al., 1995).

6.1.2 ENTROPY AND REGULARITY

Depending on the choice of entropy and divergence functionals we obtain a range of diverse estimators. For instance, if we were to choose the *unnormalized* entropy instead of the entropy, we would obtain algorithms more akin to boosting. We may also use Csiszar and Bregmann divergences. The key point is that our reasoning of estimating μ_{XY} based on an aggregate of samples with unknown labels but known label proportions is still applicable.

6.2 Unknown Test Label Proportions

In many practical applications we may not actually know the label proportions on the test set. For instance, when deploying the algorithm to assess the spam in a user's mailbox we will *not* know

what the fraction would be. Nor is it likely that the user would be willing or able or trustworthy enough to provide a reliable estimate. This means that we need to estimate those proportions in addition to the class means μ_x^{class}.

We may use a fairly straightforward simplification of the covariate shift correction procedure of Huang et al. (2007) in this context. The basic idea is to exploit the fact that there the map $p(x) \to \mu[p(x)] = \mathbf{E}_x[\psi(x)]$ is injective for characteristic kernels (Sriperumbudur et al., 2008). Examples of such a characteristic kernel is Gaussian RBF, Laplacian, and B_{2n+1}-splines. This means that as long as the conditional distributions $p(x|y)$ are different for different choices of y we will be able to recover the test label proportions by the simple procedure of minimizing the distance between $\mu[p]$ and $\sum_y \alpha_y \mu[p(x|y)]$. While we may not have access to the true expectations we are still able to estimate $\mu_x^{\text{class}}[y]$ for all $y \in \mathcal{Y}$. This leads to the optimization problem

$$\begin{aligned} &\underset{\alpha}{\text{minimize}} \quad \left\| \frac{1}{m} \sum_{i=1}^{m} \psi(x_i) - \sum_{y \in \mathcal{Y}} \alpha_y \mu_X^{\text{class}}[y] \right\|^2 \\ &\text{subject to } \alpha_y \geq 0 \text{ and } \sum_{y \in \mathcal{Y}} \alpha_y = 1. \end{aligned} \tag{15}$$

Here the sum is taken over the elements of the test set, that is $x_j \in X$. Very similar bounds to those by Huang et al. (2007) can be obtained and they are omitted for the sake of brevity as the reasoning is essentially identical.

Note that obviously (15) may be used *separately* from the previous discussion, that is, when the training proportions are known but the test proportions are not. However, we believe that the most significant benefit is obtained in using both methods in conjunction since many practical situations exhibit both problems simultaneously.

7. Related Work and Alternatives

While being highly relevant in practice, the problem has not seen as much attention by researchers as one would expect. Some of the few works which cover a related subject are those by Chen et al. (2006) and Musicant et al. (2007), and by Kück and de Freitas (2005). We hope that our work will stimulate research in this area as relevant problems are fairly widespread.

7.1 Transduction

In transduction one attempts to solve a related problem: the patterns x_i on the test set are known, usually also some label proportions on the test set are known but obviously the actual labels on the test set are *not* known. One way of tackling this problem is to perform transduction by enforcing a proportionality constraint on the unlabeled data, e.g. via a Gaussian Process model (Gärtner et al., 2006; Mann and McCallum, 2007).

At first glance these methods might seem applicable for our problem but they do require that we have at least some labeled instances of *all classes* at our disposition which need to be drawn in an unbiased fashion. This is clearly not the case in our setting. That said, it is well possible to use our setting in the context of transduction, that is, to replace the unknown mean μ_{XY}^{test} on the test set by the empirical estimate on the training set. Such strategies lead to satisfactory performance on par with (albeit not exceeding) existing transduction approaches.

7.2 Self Consistent Proportions

Kück and de Freitas (2005) introduced a more informative variant of the binary multiple-instance learning, in which groups of instances are given along with estimates of the fraction of positively-labeled instances per group. The authors build a fully generative model of the process which determines the assignment of observations to individual bags. Such a procedure is likely to perform well when a large number of bags is present.

In order to deal with the estimation of the missing variables a MCMC sampling procedure is used. While Kück and de Freitas (2005) describe the approach only for a binary problem, it could be extended easily to multiclass settings.

In a similar vein, Chen et al. (2006) and Musicant et al. (2007) also use a self-consistent approach where the conditional class estimates need to match the observed ones. Consequently it shares the same similar drawbacks, since we typically only have as many sets as classes.

7.3 Conditional Probabilities

A seemingly valid alternative approach is to try building a classifier for $p(i|x)$ and subsequently recalibrating the probabilities to obtain $p(y|x)$, e.g. via $p(y|i)$. At first sight this may appear promising since this method is easily implemented by most discriminative methods. The idea would be to reconstruct $p(y|x)$ by

$$p(y|x) = \sum_i \pi_{iy} p(i|x).$$

However, this is not a useful estimator in our setting for a simple reason: it assumes the conditional independence $y \perp\!\!\!\perp x \mid i$, which obviously does not hold. Instead, we have the property that $i \perp\!\!\!\perp x| y$, that is, the distribution over x for a given class label does not depend on the bag. This mismatch in the probabilistic model can lead to disastrous estimates as the following simple example illustrates:

Example 1 *Assume that $\mathcal{X}, \mathcal{Y} = \{1,2\}$ and that $p(y=1|x=1) = p(y=2|x=2) = 1$. In other words, the estimation problem is solvable since the classes are well separated. Moreover, assume that π is given by*

$$\pi = \begin{bmatrix} 0.5-\varepsilon & 0.5+\varepsilon \\ 0.5 & 0.5 \end{bmatrix} \text{ for } 0 < \varepsilon \ll 1.$$

Here, $p(i|x)$ is useless for estimating $p(y|x)$, since we will only exceed random guessing by at most ε. On the other hand, it is easily possible to obtain a good estimate for μ_{XY} by our proposed procedure.

The reason for this failure can be found in the following expansion

$$p(y|x) = \sum_i p(y|x,i)p(i|x) \neq \sum_i p(y|i)p(i|x) \text{ since } p(y|x,i) \neq p(y|i). \tag{16}$$

The problem with (16) is that the estimator does not really attempt to compute the probability $p(y|x)$, which we are interested in but instead, it attempts to discern which mixture distribution p_i the observation x most likely originated from. For this to work we would need good probability estimates as the *basis* of reweighting. Our approach tackles the problem at the source by recalibrating the sufficient statistics directly.

7.4 Reduction to Binary

For binary classification and real-valued classification scores we may resort to a rather straightforward heuristic: build a classifier which is able to distinguish between the sets X_1 and X_2 and subsequently threshold labels such that the appropriate fraction of observations in X_1 and X_2 matches the proper labels. The intuition is that since the bags X_1 and X_2 do contain some information about how the two classes differ, we should be able to use this information to distinguish between different class labels.

It is likely that one might be able to obtain a proper reduction bound in this context. However, extensions to multi-class are highly nontrivial. It also turns out that even in the binary case this method, while overall fairly competitive, is inferior to our approach.

7.5 Density Estimation

One way of obtaining $p(x|i)$ is to carry out density estimation. While, in principle, this approach is flawed because of the incorrect conditional independence assumptions, it can still lead to acceptable results whenever each of the bags contains one majority class. This allows us to obtain

$$p(x|y) = \sum_i \left[\pi^{-1}\right]_{yi} p(x|i).$$

To re-calibrate the probability estimates Bayes' theorem is invoked to compute posterior probabilities. Since this approach involves density estimation it tends to fail fairly catastrophically for high-dimensional data due to the curse of dimensionality. These problems are also manifest in the experiments.

8. Experiments

Data Sets: We use binary and three-class classification data sets from the UCI repository[1] and the LibSVM site.[2] If separate training and test sets are available, we merge them before performing nested 10-fold cross-validation. Since we need to generate as many splits as classes, we limit ourselves to three classes.

For the binary data sets we use half of the data for X_1 and the rest for X_2. We also remove all instances of class 2 from X_1. That is, the conditional class probabilities in X_2 match those from the repository, whereas in X_1 their counterparts are deleted.

For three-class data sets we investigate two different partitions. In scenario A we use class 1 exclusively in X_1, class 2 exclusively in X_2, and a mix of all three classes weighted by $(0.5 \cdot p(1), 0.6 \cdot p(2), 0.7 \cdot p(3))$ to generate X_3. In scenario B we use the following splits

$$\begin{bmatrix} c_1 \cdot 0.4 \cdot p(1) & c_1 \cdot 0.2 \cdot p(2) & c_1 \cdot 0.2 \cdot p(3) \\ c_2 \cdot 0.1 \cdot p(1) & c_2 \cdot 0.2 \cdot p(2) & c_2 \cdot 0.1 \cdot p(3) \\ c_3 \cdot 0.5 \cdot p(1) & c_3 \cdot 0.6 \cdot p(2) & c_3 \cdot 0.7 \cdot p(3) \end{bmatrix}.$$

Here the constants c_1, c_2 and c_3 are chosen such that the probabilities are properly normalized. As before, X_3 contains half of the data.

1. UCI can be found at `http://archive.ics.uci.edu/ml/`.
2. LibSVM can be found at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/`.

Model Selection: As stated, we carry out a *nested* 10-fold cross-validation procedure: 10-fold cross-validation to assess the performance of the estimators; within each fold, 10-fold cross-validation is performed to find a suitable value for the parameters.

For supervised classification, i.e. discriminative sorting, such a procedure is quite straightforward because we can directly optimize for classification error. For kernel density estimation (KDE), we use the log-likelihood as our criterion.

Due to the high number of hyper-parameters (at least 8) in MCMC, it is difficult to perform *nested* 10-fold cross-validation. Instead, we choose the *best* parameters from a simple 10-fold crossvalidation run. In other words, we are giving the MCMC method an unfair advantage over our approach by reporting the best performance during the model selection procedure.

Finally, for the re-calibrated sufficient statistics $\hat{\mu}_{XY}$ we use the estimate of the log-likelihood on the validation set as the criterion for cross-validation, since no other quantity, such as classification errors is readily available for estimation.

Algorithms: For discriminative sorting we use an SVM with a Gaussian RBF kernel whose width is set to the median distance between observations (Schölkopf, 1997); the regularization parameter is chosen by cross-validation. The same strategy applies for our algorithm. For KDE, we use Gaussian kernels. Cross-validation is performed over the kernel width. For MCMC, 10 000 samples are generated after a burn-in period of 10 000 steps (Kück and de Freitas, 2005).

Optimization: Bundle methods (Smola et al., 2007; Teo et al., 2007) are used to solve the optimization problem in Algorithm 1. For our regularized log-likelihood, the solver converges to ε precision in $O(\log(1/\varepsilon))$ steps.

Results: The experimental results are summarized in Table 3. Our method outperforms KDE and discriminative sorting. In terms of computation, our approach is somewhat more efficient, since it only needs to deal with a smaller sample size (only X rather than the union of all X_i). The training time for our method is less than 2 minutes for all cases, whereas MCMC on average takes 15 minutes and maybe even much longer when the number of active kernels and/or observations are high. Note that KDE fails on two data sets due to numerical problems (high dimensional data).

Our method also performs well on multiclass data sets. As described in Section 5.2, the quality of our minimizer of the negative log-posterior depends on the mixing matrix and this is noticeable in the reduction of performance for the dense mixing matrix (scenario B) in comparison to the better conditioned sparse mixing matrix (scenario A). In other words, for ill conditioned π even our method has its limits, simply due to numerical considerations of effective sample size.

Unknown test label proportions: In this experiment, we use binary and three-class classification data sets with the same split procedure as in the previous experiment but we select testing examples by a biased procedure to introduce unknown test label proportions. To describe our biased procedure, consider a random variable ξ_i for each point in the pool of possible testing samples where $\xi_i = 1$ means the i-th sample is being included and $\xi_i = 0$ means the sample is discarded. In our case, the biased procedure only depends on the label y, i.e. $P(\xi = 1|y = 1) = 0.5$ and $P(\xi = 1|y = -1) = 1.0$ for binary problems and $P(\xi = 1|y = 1) = 0.6$, $P(\xi = 1|y = 2) = 0.3$, and $P(\xi = 1|y = 3) = 0.1$ for three-class problems. We then estimate the test proportion by solving the quadratic program in (15) with interior point methods (or any other successive optimization procedure). Since we are interested particularly to assess the effectiveness of our test proportion estimation method, in solving (15) we assume that we can compute $\mu_X^{\text{class}}[y]$ directly, i.e. the instances are labeled. The mean square error rates of test proportions for several binary and three-class data

Data	MM	KDE	DS	MCMC	BA
ionosphere	18.4±3.2	**17.5±3.2**	**12.2±2.6**	18.0±2.1	35.8
iris	**10.0±3.6**	**16.8±3.4**	**15.4±1.1**	**21.1±3.6**	29.9
optdigits	1.8±0.5	**0.7±0.4**	9.8±1.2	2.0±0.4	49.1
pageblock	**3.8±2.3**	7.1±2.8	18.5±5.6	**5.4±2.8**	43.9
pima	**27.5±3.0**	34.8±0.6	34.4±1.7	**23.8±1.8**	34.8
tic	31.0±1.5	34.6±0.5	**26.1±1.5**	31.3±2.5	34.6
yeast	**9.3±1.5**	**6.5±1.3**	25.6±3.6	10.4±1.9	39.9
wine	**7.4±3.0**	**12.1±4.4**	18.8±6.4	**8.7±2.9**	40.3
wdbc	**7.8±1.3**	**5.9±1.2**	10.1±2.1	15.5±1.3	37.2
sonar	**24.2±3.5**	35.2±3.5	31.4±4.0	39.8±2.8	44.5
heart	**30.0±4.0**	38.1±3.8	**28.4±2.8**	**33.7±4.7**	44.9
breastcancer	**5.3±0.8**	14.2±1.6	**3.5±1.3**	**4.8±2.0**	34.5
australian	**17.0±1.7**	33.8±2.5	**15.8±2.9**	30.8±1.8	44.4
svmguide3	**20.4±0.9**	27.2±1.3	25.5±1.5	24.2±0.8	23.7
adult	**18.9±1.2**	24.5±1.3	22.1±1.4	**18.7±1.2**	24.6
cleveland	**19.1±3.6**	35.9±4.5	**23.4±2.9**	**24.3±3.1**	22.7
derm	**4.9±1.4**	27.4±2.6	**4.7±1.9**	14.2±2.8	30.5
musk	**25.1±2.3**	28.7±2.6	**22.2±1.8**	**19.6±2.8**	43.5
german	**32.4±1.8**	41.6±2.9	37.6±1.9	**32.0±0.6**	32.0
covertype	37.1±2.5	41.9±1.7	**32.4±1.8**	41.1±2.2	45.9
splice	**25.2±2.0**	35.5±1.5	**26.6±1.7**	28.8±1.6	48.4
gisette	**10.3±0.9**	†	**12.2±0.8**	50.0±0.0	50.0
madelon	**44.1±1.5**	†	**46.0±2.0**	49.6±0.2	50.0
cmc	**37.5±1.4**	43.8±0.7	45.1±2.3	46.9±2.6	49.9
bupa	**48.5±2.9**	50.8±5.1	**40.3±4.9**	50.4±0.8	49.7
protein A	**43.3±0.4**	48.9±0.9	N/A	65.5±1.7	60.6
protein B	**46.9±0.3**	55.2±1.5	N/A	66.1±2.1	60.6
dna A	**14.8±1.2**	28.1±0.6	N/A	39.8±2.6	41.6
dna B	**31.3±1.3**	**30.4±0.7**	N/A	41.5±0.1	41.6
senseit A	**19.8±0.1**	44.2±0.0	N/A	‡	44.2
senseit B	**21.1±0.1**	44.2±0.0	N/A	‡	44.2

Table 3: Classification error on UCI/LibSVM data sets. Errors are reported in mean ± standard error. The best result and those not significantly worse than it, are highlighted in boldface. We use a one-sided paired t-test with 95% confidence. `MM`: Mean Map (our method); `KDE`: Kernel Density Estimation; `DS`: Discriminative Sorting (only applicable for binary classification); `MCMC`: the sampling method; `BA`: Baseline, obtained by predicting the major class. †: Program fails (too high dimensional data - only `KDE`). ‡: Program fails (large data sets - only `MCMC`).

sets are presented in Table 4. The results show that our proportion estimation method works reasonably well.

Overdetermined systems: Here we are interested to assess the performance of our estimator with optimized weights when we have more data sets n than class labels $|\mathcal{Y}|$ with varying number of observations m_i per data set. We simulate the problem in binary settings with the following split ($n = 8$)

$$\begin{bmatrix} c_1 \cdot 0.25 \cdot p(1) & c_1 \cdot 0.10 \cdot p(2) \\ c_2 \cdot 0.15 \cdot p(1) & c_2 \cdot 0.10 \cdot p(2) \\ c_3 \cdot 0.05 \cdot p(1) & c_3 \cdot 0.20 \cdot p(2) \\ c_4 \cdot 0.05 \cdot p(1) & c_4 \cdot 0.10 \cdot p(2) \\ c_5 \cdot 0.05 \cdot p(1) & c_5 \cdot 0.00 \cdot p(2) \\ c_6 \cdot 0.05 \cdot p(1) & c_6 \cdot 0.05 \cdot p(2) \\ c_7 \cdot 0.05 \cdot p(1) & c_7 \cdot 0.15 \cdot p(2) \\ c_8 \cdot 0.35 \cdot p(1) & c_8 \cdot 0.30 \cdot p(2) \end{bmatrix}$$

and the split ($n = 6$) in three-class settings is as follows

$$\begin{bmatrix} c_1 \cdot 0.30 \cdot p(1) & c_1 \cdot 0.10 \cdot p(2) & c_1 \cdot 0.00 \cdot p(3) \\ c_2 \cdot 0.10 \cdot p(1) & c_2 \cdot 0.10 \cdot p(2) & c_2 \cdot 0.20 \cdot p(3) \\ c_3 \cdot 0.05 \cdot p(1) & c_3 \cdot 0.00 \cdot p(2) & c_3 \cdot 0.05 \cdot p(3) \\ c_4 \cdot 0.05 \cdot p(1) & c_4 \cdot 0.20 \cdot p(2) & c_4 \cdot 0.05 \cdot p(3) \\ c_5 \cdot 0.00 \cdot p(1) & c_5 \cdot 0.05 \cdot p(2) & c_5 \cdot 0.10 \cdot p(3) \\ c_6 \cdot 0.50 \cdot p(1) & c_6 \cdot 0.55 \cdot p(2) & c_6 \cdot 0.60 \cdot p(3) \end{bmatrix}.$$

We use BFGS to obtain the optimal weights of the minimization problem in (11). We perform 10-fold cross validation with respect to the log-likelihood. The error rates are presented in Table 5. For all cases except one, the estimator with optimized weights improves error rates compared with the unweighted one.

Binary data sets

Data	**MSE**
australian	0.00804±0.00275
breastcancer	0.00137±0.00063
adult	0.00610±0.00267
derm	0.00398±0.00175
gisette	0.00331±0.00108
wdbc	0.00319±0.00103

Three-class data sets

Data	**MSE**
protein	0.00290±0.00066
dna	0.00339±0.00075
senseit	0.00072±0.00031

Table 4: Unknown test label proportion case. Square errors of estimating the test proportions on UCI/LibSVM data sets. The 10-run errors are reported in mean ± standard error.

Stability of Mixing Matrices: Lastly, we are interested to assess the performance of our proposed method when the given mixing matrix π are perturbed so that they do not exactly match how the data is generated. We used binary classification data sets and defined the perturbed mixing

Binary data sets

Data	unweighted	weighted
wdbc	23.29±2.68	**14.22±1.79**
australian	34.44±4.03	29.58±3.71
svmguide3	24.28±2.20	**18.50±1.73**
gisette	8.77±1.05	7.69±0.51
splice	33.43±1.65	**21.12±2.59**

Three-class data sets

Data	unweighted	weighted
protein	57.46±0.02	57.46±0.02
senseit	28.25±2.60	23.51±0.78
dna	20.01±1.26	**16.80±1.19**

Table 5: Overdetermined systems. Errors of weighted/unweighted estimators for overdetermined systems on UCI/LibSVM data sets. The 10-fold cross validation errors are reported in mean ± standard error. The numbers in boldface are significant with 95% confidence (one-sided paired t-test).

matrix as

$$\tilde{\pi} = \pi + \Delta = \begin{bmatrix} 1 & 0 \\ \rho & 1-\rho \end{bmatrix} + \begin{bmatrix} -\varepsilon_1 & \varepsilon_1 \\ \varepsilon_2 & -\varepsilon_2 \end{bmatrix}.$$

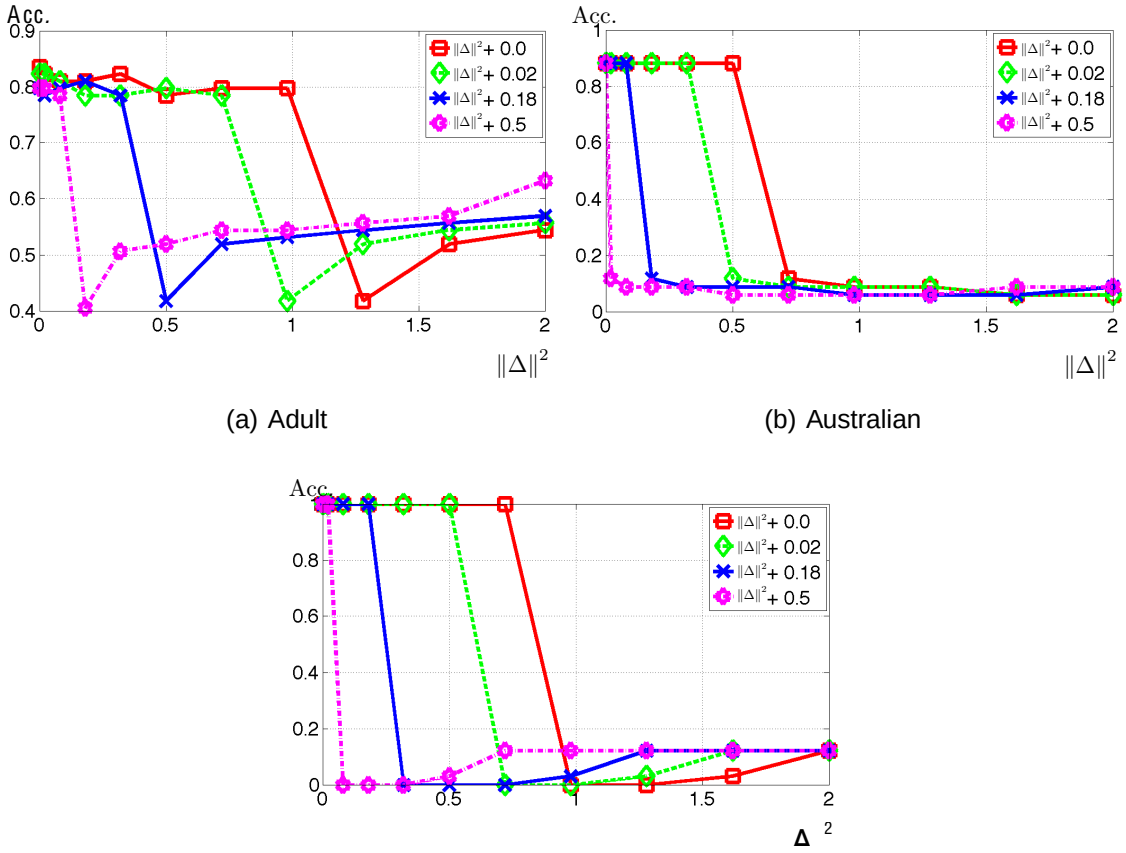
We varied $\varepsilon_1 \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $\varepsilon_2 \in \{0.0, 0.1, 0.3, 0.5\}$ and measured the performance as a function of the size of the perturbation, $\eta = \|\Delta\|^2 = tr(\Delta^\top \Delta)$. Note that unperturbed mixing matrix refer to the case of $\{\varepsilon_1, \varepsilon_2\} = \{0, 0\}$. The experiments are summarized in Figure 2. The results suggest that for a reasonable size of perturbations, our method is stable.

9. Conclusion

In this paper we obtained a rather surprising result, namely that it is possible to consistently reconstruct the labels of a data set if we can only obtain information about the proportions of occurrence of each class (in at least as many data aggregates as there are classes). In particular, we proved that up to constants, our algorithm enjoys the same rates of convergence afforded to methods which have full access to all label information.

This finding has significant implications with regard to the amount of privacy afforded by summary statistics. In particular, it implies that whenever accurate summary statistics exist and whenever the available individual statistics are highly dependent on the summarized random variable we will be able to perform inference on the summarized variable with a high degree of confidence. In other words, some techniques used to anonymize observations, e.g. demographic data, may not be really safe (at least when it is possible to estimate the missing information, provided enough data).

Recently Chiaia et al. (2007) applied a summarization technique to infer drug use based on the concentration of metabolites in the sewage of cities, suburbs or at an even more finely grained resolution. While this only provides aggregate information about the proportions of drug users, such data, in combination with detailed demographic information might be used to perform more detailed inference with regard to the propensity of individuals to use controlled substances. It is in these types of problem where our method could be applied straightforwardly.



(a) Adult (b) Australian

(c) Breastcancer

Figure 2: Performance accuracy of binary classification data sets ($n = |\mathcal{Y}| = 2$) as a function of the amount of perturbation applied to the mixing matrix, $\|\Delta\|^2 = tr(\Delta^\top \Delta)$ with $\Delta = \tilde{\pi} - \pi$. 2(a): Adult, 2(b): Australian and 2(c): Breastcancer data sets. x-axis denotes $\|\Delta\|^2$ as a function of $\varepsilon_1 \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Color coded plots denote $\|\Delta\|^2$ as a function of $\varepsilon_2 \in \{0.0, 0.1, 0.3, 0.5\}$, for example red colored plot refers to performance when only label proportions of the first set are perturbed.

Acknowledgments

We thank Hendrik Kück and Choon Hui Teo for providing us with optimization code. NICTA is funded by the Australian Government's Backing Australia's Ability and the Centre of Excellence programs. This work was supported by the PASCAL network of excellence of the European Union.

References

Y. Altun and A.J. Smola. Unifying divergence minimization and statistical inference via convex duality. In H.U. Simon and G. Lugosi, editors, *Proc. Annual Conf. Computational Learning Theory*, LNCS, pages 139–153. Springer, 2006.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, 2005.

B.C. Chen, L. Chen, R. Ramakrishnan, and D.R. Musicant. Learning from aggregate views. In L. Liu, A. Reuter, K.Y. Whang, and J. Zhang, editors, *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, pages 3–12, Atlanta, GA, 2006.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, May 1995.

A. C. Chiaia, C. Banta-Green, L. Power, D. L. Sudakin, and J. A. Field. Community burdens of methamphetamine and other illicit drugs. In *6th International Conference on Pharmaceuticals and Enocrine Disrupting Chemicals in Water*, 2007.

M. Dudík and R. E. Schapire. Maximum entropy distribution estimation with generalized regularization. In Gábor Lugosi and Hans U. Simon, editors, *Proc. Annual Conf. Computational Learning Theory*. Springer Verlag, June 2006.

T. Gärtner, Q.V. Le, S. Burton, A.J. Smola, and S.V.N. Vishwanathan. Large-scale multiclass transduction. In *Neural Information Processing Systems*, pages 411–418. MIT Press, 2006.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. Technical Report 156, Max-Planck-Institut für biologische Kybernetik, 2006. To appear in the Annals of Statistics.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.

H. Kück and N. de Freitas. Learning about individuals from group statistics. In *Uncertainty in Artificial Intelligence (UAI)*, pages 332–339, Arlington, Virginia, 2005. AUAI Press.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

G. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In Zoubin Ghahramani, editor, *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007),Corvallis, OR*, pages 593–600. Omnipress, 2007.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Trans. Inform. Theory*, 48(1):251–263, 2002.

D.R. Musicant, J. Christensen, and J.F. Olson. Supervised learning by training on aggregate outputs. In *IEEE International Conference on Data Mining*, 2007.

N. Quadrianto, A. Smola, T. Caetano, and Q. Le. Estimating labels from label proportions. In W. Cohen, A. McCallum, and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 776–783. Omnipress, 2008.

B. Schölkopf. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997. Download: http://www.kernel-machines.org.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

A.J. Smola, S. V. N. Vishwanathan, and Q.V. Le. Bundle methods for machine learning. In Daphne Koller and Yoram Singer, editors, *Advances in Neural Information Processing Systems 20*, Cambridge MA, 2007. MIT Press.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 111–122, 2008.

C.H. Teo, Q. Le, A.J. Smola, and S.V.N. Vishwanathan. A scalable modular convex solver for regularized risk minimization. In *Proc. ACM Conf. Knowledge Discovery and Data Mining (KDD)*. ACM, 2007.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, 2005.

P.Å. Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2), 1973.